Data Drift

As we mentioned earlier, **data drift** is change in the input data for a model. Over time, data drift causes degradation in the model's performance, as the input data drifts farther and farther from the data on which the model was trained.

QUESTION 1 OF 2

Below are the different *causes of data drift* that we just discussed. Can you match each of them with the correct example?

Submit to check your answer choices!

EXAMPLE	CAUSE OF DATA DRIFT
A change in customer behavior over time.	Natural drift in the data
A sensor breaks and starts providing inaccurate readings.	Data quality issues
Two features that used to be correlated are no longer correlated.	Covariate shift / Change in relationship between features
A sensor is replaced, causing the units of measurement to change (e.g., from minutes to seconds).	Upstream process changes

SUBMIT

Monitoring for Data Drift

As we noted, data drift is one of the main reasons that model performance gets worse over time. Fortunately, Azure Machine Learning allows you to set up *dataset monitors* that can alert you about data drift and even take automatic actions to correct data drift.

Remember, the process of monitoring for data drift involves:

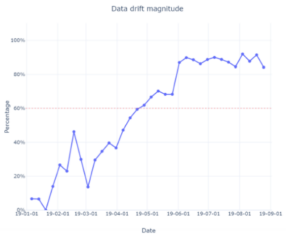
- Specifying a **baseline dataset** – usually the training dataset
- Specifying a **target dataset** – usually the input data for the model
- Comparing these two datasets over time, to monitor for differences

Here are a couple different types of comparisons you might want to make when monitoring for data drift:

- **Comparing input data vs. training data.** This is a proxy for model accuracy; that is, an increased difference between the input vs. training data is likely to result in a decrease in model accuracy.
- **Comparing different samples of time series data.** In this case, you are checking for a difference between one time period and another. For example, a model trained on data collected during one season may perform differently when given data from another time of year. Detecting this seasonal drift in the data will alert you to potential issues with your model's accuracy.

QUESTION 2 OF 2

Here's the graph of data-drift magnitude that we looked at in the video:



In this example, a baseline January dataset was compared with a dataset containing all 2019 data. Which of these statements about the graph are true?

(Select all that apply.)

- ☐ The baseline vs. target datasets show a lot of difference in January
- ☑ The baseline vs. target datasets show very little difference in January
- ☑ The baseline vs. target datasets show a lot of difference in August
- ☐ The baseline vs. target datasets show very little difference in August

SUBMIT

NEXT